EXTENDING OUTPUT ATTENTIONS IN RECURRENT NEURAL NETWORKS FOR DIALOG GENERATION

Chanseung Lee
Lane College
Eugene, OR USA

ABSTRACT

Attention mechanism in recurrent neural networks has been widely used in natural language processing. In this paper, the research team explore a new mechanism of extending output attention in recurrent neural networks for dialog systems. The new attention method was compared with the current method in generating dialog sentence using a real dataset. Our architecture exhibits several attractive properties such as better handle long sequences and, it could generate more reasonable replies in many cases.

KEYWORDS

Deep learning; Dialog Generation; Recurrent Neural Networks; Attention

1. INTRODUCTION

In the domain of natural language dialogue generation, conventional statistical approaches tend to rely extensively on hand-crafted rules and templates which prevent training on the large human conversational corpora that are becoming increasingly available. Therefore, a great deal of attention has been recently paid to developing data driven methods for generating novel natural language responses.

Amongst data driven language generation models, recurrent neural networks (RNN), long short-term memory [1] and gated recurrent [2] neural networks in particular, have been widely used in dialog systems [3] and machine translation [4]. In RNN-based dialog generative models, the meaning of a sentence is mapped into a fixed-length vector representation and then they generate a translation based on that vector. By not relying on things like n-gram counts and instead trying to capture the higher-level meaning of a text, the systems generalize to new sentences better than many other approaches.

One of the recent advancement in sequence learning is attention mechanism. With an attention mechanism, we no longer try encode the full source sentence into a fixed-length vector. Rather, researchers allow the decoder to ‘attend’ to different parts of the source sentence at each step of the output generation. Neural attention has proven very successful for various sequence-to-sequence tasks by associating salient items in the source sequence with the generated item in the target sequence [5] [6].
In this paper, the research team introduce a new attention method, called ‘output attention’, in a recurrent neural network (RNN) language model. At every time step of a RNN, a weighted average of all the previous outputs will be used as an extra input to the function that computes the next output state. The basic idea of output attention is that when you generate next word, the previous words you already generated are clearly important in choosing next word. Therefore, researchers assign different weights to each of the previous output and their combined attention value is fed into the current output function. With output attention mechanism, the network can take the outputs produced many time steps earlier into consideration.

2. RELATED WORK

Many work has been done in the area of statistical machine translation-based response generation [1] [7] [8] [9]. Ritter, Cherry, and Dolan [10] formulate dialogue response generation as a statistical phrase-based machine translation problem, which requires no explicit hand-crafted rules. The recent success of RNNs in statistical machine translation [1] [7] has inspired the application of such models to the field of dialogue modeling.

Other researchers have recently used SEQ2SEQ to directly generate responses in an end-to-end fashion without relying on SMT phrase tables [11] [12]. Vinyals and Le [11] and Shang, Lu, and Li [12] employ an RNN to generate responses in human-to-human conversations by treating the conversation history as one single temporally ordered sequence. In such models, the distant relevant context in the history is difficult to recall. Some efforts have been made to overcome this limitation.

Sordoni et al. [9] separately encoded the most recent message and all the previous context using a bag-of-words representation, which is decoded using an RNN. This approach equates the distance of each word in the generated output to all the words in the conversation history, but loses the temporal information of the history.

Serban et al. [2] design a hierarchical model that stacks an utterance-level RNN on a token-level RNN, where the utterance-level RNN reduces the number of computational steps between utterances. Wen et al. [13] and Wen et al. [14] improve spoken dialog systems via multi-domain and semantically conditioned neural networks on dialog act representations and explicit slot-value formulations.

Recently, Tran, Bisazza, and Monz [15] demonstrated that the memory network mechanism can improve the effectiveness of the neural language model.

In this paper, the team of researchers propose a new attention-based neural language model for dialogue modeling that learns how a conversation evolves as a whole, rather than only how the most recent response is generated.

3. MODEL ARCHITECTURE

In this section, researchers describe the architecture of the new extended output attention model
and discuss how these additional attentions can help to achieve efficient dialog generation for sequence-to-sequence learning.

**Encoder:** A recurrent neural network is a neural network that consists of a hidden state \( h \) and an optional output \( y \) which operates on a variable-length sequence \( X = \{x_1, x_n\} \). At each time step \( t \), the hidden state \( h_t \) of RNN is updated by

\[
    h_t = f(h_{t-1}, x_t)
\]  

Where \( f \) is a non-linear activation function. \( f \) is the nonlinear function in the recurrent unit, which can be implemented in a non-linear activation function, or Long Short-Term Memory (LSTM) [3], or Gated Recurrent Unit (GRU) [4].

**Decoder:** In traditional model architecture, define each conditional probability as follows

\[
    p(y_t | y_1, \cdots, y_{t-1}, X) = g(y_{t-1}, s_t, c_t)
\]  

where \( s_t \) is an RNN hidden state for time \( t \), computed by

\[
    s_t = f(s_{t-1}, y_{t-1}, c_t)
\]  

The probability is conditioned on a traditional attention vector \( c_t \) for each target word \( y_t \). This implements a mechanism of attention in the decoder.

The decoder decides parts of the source sentence to pay attention to. By letting the decoder have an attention mechanism, the researcher relieves the encoder from the burden of having to encode all information in the source sentence into a fixed length vector.

The context vector \( c_t \) depends on a sequence of annotations \( (h_1, \ldots, h_t) \) to which an encoder maps the input sentence. Each annotation \( h_i \) contains information about the whole input sequence with a strong focus on the parts surrounding the \( i \)-th word of the input sequence.

The context vector \( c_t \) is, then, computed as a weighted sum of these annotations \( h_i \):

\[
    c_t = \sum_{j=1}^{t} \alpha_{tj} \, h_j
\]  

The weight \( \alpha_{tj} \) of each annotation \( h_j \) is computed by

\[
    \alpha_{tj} = \frac{\exp(e_{tj})}{\sum_k \exp(e_{tk})}
\]
Where,

$$e_{tj} = a(h_j, h_t)$$  \hspace{1cm} (6)

represents a score of how well the inputs around position \(j\) and the output at position \(t\) match. The probability \(a\) reflects the importance of the annotation \(h_i\) with respect to the previous hidden state \(s_{t-1}\) in deciding the next state \(s_t\) and generating \(y_i\). Figure 1 shows the structure of traditional attention method.

![Figure 1 Architecture of traditional attention](image)

In this paper, the team of researchers modify the recurrence formula by adding a new output attention vector \(d_t\) to the input of the LSTM. As for the following conditional probability, Eq. (2), the term \(s_t\) in Eq. (3) is now modified as follows

$$s_t = f(s_{t-1}, d_{t-1}, c_t)$$  \hspace{1cm} (7)

$$d_t = \sum_{j=1}^{t} \beta_{tj} y_j$$  \hspace{1cm} (8)

$$\beta_{tj} = \frac{\exp(f_{tj})}{\sum_k \exp(f_{tk})}$$  \hspace{1cm} (9)
The probability $\beta_i$ reflects the importance of the annotation $h_i$ with respect to the previous hidden state $s_{i-1}$ in deciding the next state $s_i$ and generating $y_i$.

$$f_{tj} = b(s_j, h_t) \quad (10)$$

The definition of $c_t, \alpha_{ij}, e_{ij}$ remain unchanged. Figure 2 shows the structure of the proposed model in this paper.

![Architecture of extended output attention](image)

In Figure 2, $d_t$ represents *output attention* vector, and is a new extended attention vector for output values.

**4. EXPERIMENTAL RESULTS**

In this section, researchers describe the experimental results with some datasets and show some samples of the interactions with the system that we trained. They used Movie Dialog Data for the experiment. This dataset contains a metadata-rich collection of fictional conversations extracted from raw movie scripts. It contains 220,579 conversational exchanges between 10,292 pairs of movie characters and total of from 617 movies total 304,713 utterances from 9,035 characters.

Researchers trained a single layer LSTM with 1024 memory cells using stochastic gradient descent with gradient clipping. The vocabulary consists of the most common 20K words, which includes special tokens indicating turn taking and actor. Below is a few samples of
simulated dialog sessions. Table 1 shows some sample utterance between machine and human. In Table 1, ‘Input’ is given by human and ‘Output attention’ is the response generated by the proposed system in this paper and ‘Attention’ is the response from traditional attention method.

Researchers find it encouraging that the model can better understand contexts, and generate more reasonable sentences compared to the traditional model. The model does so without any explicit knowledge representation component except for the additional output attention.

One drawback of this basic model is that it only gives simple, short, and sometimes unsatisfying answers to our questions as can be seen in Table 1. Indeed, if a researcher asks not identical but semantically similar questions, the answers can sometimes be inconsistent. This is expected due to the simplicity of our model and the dataset in our experiments.

Table 1 Summary of dialog

<table>
<thead>
<tr>
<th>Input</th>
<th>Output attention</th>
<th>Attention</th>
</tr>
</thead>
<tbody>
<tr>
<td>How do you do?</td>
<td>i am always bad</td>
<td>i am not this</td>
</tr>
<tr>
<td>are you religious ?</td>
<td>we're all like that</td>
<td>we're all here for you</td>
</tr>
<tr>
<td>i'm thirty four .</td>
<td>i'm twenty years old. how old are you ?</td>
<td>i'm twenty years old.</td>
</tr>
<tr>
<td>all right i never should have bought them !</td>
<td>they're killing me !</td>
<td>they're killing me !</td>
</tr>
<tr>
<td>what's so interesting about him ?</td>
<td>i'm just saying that</td>
<td>i'm just on vacation</td>
</tr>
<tr>
<td>i can't blame anybody for anything i do</td>
<td>we're partners</td>
<td>we're gonna ?</td>
</tr>
<tr>
<td>much more terry. much much more !</td>
<td>i'm trying my best</td>
<td>i'm trying to find out with</td>
</tr>
<tr>
<td>they said she was here</td>
<td>she's not at home ?</td>
<td>she's not a ? she's just</td>
</tr>
<tr>
<td>sure you can if you want to.</td>
<td>i'm not sure i can</td>
<td>i'm not sure i can i . to</td>
</tr>
<tr>
<td>bank teller is just a start. i got plans.</td>
<td>you're talking about giving up baseball ?</td>
<td>you're talking about a ?</td>
</tr>
<tr>
<td>i am not smiling</td>
<td>i am always down</td>
<td>i am not a loser</td>
</tr>
<tr>
<td>who's leaving?</td>
<td>we're leaving</td>
<td>we leaving</td>
</tr>
<tr>
<td>are you alright ?</td>
<td>i am good. thanks how are you</td>
<td>i am good</td>
</tr>
<tr>
<td>i'm not even sure why</td>
<td>i'm not sure</td>
<td>i'm so honey</td>
</tr>
</tbody>
</table>

5. CONCLUSIONS

In this paper, researchers propose a new output attention mechanism to improve the coherence of neural dialogue language models. The new attention allows each generated word to choose which related word it wants to align to in the increasing conversation history. Our modest results show that it can generate simple and more coherent conversations. Even though the dialog the model generates has obvious limitations, it is clear that another layer of
output attention allows the model to generate more meaningful responses. As future work, the model.

REFERENCES


[7] Ilya Sutskever, Oriol Vinyals, Quoc V. Le, Sequence to Sequence Learning with Neural Networks, NIPS 2014

[8] Minh-Thang Luong, Hieu Pham, Christopher D. Manning, Effective Approaches to Attention-based Neural Machine Translation, EMNLP 2015


[12] Lifeng Shang, Zhengdong Lu, Hang Li, Neural Responding Machine for Short-Text Conversation, ACL 2015
